Weaving a Semantic Web of Credibility Reviews for Explainable Misinformation Detection (Extended Abstract)

Conference or Workshop Item

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Sister Conferences Best Papers Track

Weaving a Semantic Web of Credibility Reviews for Explainable Misinformation Detection (Extended Abstract)

Ronald Denaux[1] , **Martino Mensio**[2] , **Jose Manuel Gomez-Perez**[1] and **Harith Alani**[2]
[1]Expert System Iberia
[2]Knowledge Media Institute, The Open University, UK
{rdenaux, jmgomez}@expert.ai, {martino.mensio, h.alani}@open.ac.uk

Abstract

This paper summarises work where we combined semantic web technologies with deep learning systems to obtain state-of-the art explainable misinformation detection. We proposed a conceptual and computational model to describe a wide range of misinformation detection systems based around the concepts of credibility and reviews. We described how Credibility Reviews (CRs) can be used to build networks of distributed bots that collaborate for misinformation detection which we evaluated by building a prototype based on publicly available datasets and deep learning models.

1 Introduction

Although misinformation is not a new problem, the Web –due to the pace of news cycles combined with social media, and the information bubbles it creates– has increasingly evolved into an ecosystem where misinformation can thrive [Marwick and Lewis, 2017] with negative societal effects. Tackling misinformation[1] is not something that can be achieved by a single organization –as evidenced by struggling efforts by the major social networks– as it requires decentralisation, common conceptualisations, transparency and collaboration [Cazalens *et al.*, 2018].

Technical solutions for computer-aided misinformation detection and fact-checking have recently been proposed [Babakar and Moy, 2016; Hassan *et al.*, 2017] and are essential due to the scale of the Web. The research community has defined various NLP and information retrieval tasks including check-worthiness [Nakov *et al.*, 2018] and stance detection [Schiller *et al.*, 2021; Pomerleau and Rao, 2017], while others aim to use text classification as a means of detecting deceptive language [Pérez-Rosas *et al.*, 2018] or rumours [Zubiaga *et al.*, 2018]. However, these systems only solve part of the problem and it requires custom integration. Furthermore, a lack of hand-curated data, maturity and scope of current AI systems, means assessing *veracity* [Papadopoulos *et al.*, 2016] is not feasible and often such systems do not generalize well to new types of data. Hence the value of the current systems is not so much their accuracy, but rather their capacity of retrieving potentially relevant information that can help human fact-checkers, who are the main intended users of such systems and are ultimately responsible for verifying/filtering the results such systems provide. Therefore, a main challenge is developing automated systems which can help the *general public*, and social media users in particular, to assess the credibility of web content, which requires explainable results by AI systems. This points towards the need for hybrid approaches that enable the use of the best of deep learning-based approaches, but also of symbolic knowledge graphs to enable better collaboration between large platforms, fact-checkers, the general public and other stakeholders like policy-makers, journalists, webmasters, and influencers.

Our intuition was to focus on *credibility* rather than accuracy. Credibility, defined as an estimation of factuality based on available signals or evidence, stems from earlier work MisinfoMe [Mensio and Alani, 2019b; Mensio and Alani, 2019a] which in turn borrowed from social science, media literacy and journalism research. There is also ongoing work on W3C Credibilty Signals[2], which aims to define a vocabulary to specify *credibility indicators* that may be relevant for assessing the credibility of some web content. To the best of our knowledge, this is still work in progress and no systems are implementing the proposed vocabularies.

In this paper, we summarize work presented at the International Semantic Web Conference [Denaux and Perez-Gomez, 2020], where we proposed a design on how to use semantic web technologies to aid in resolving such challenges. Our contributions are:

- a datamodel and architecture of distributed agents for composable credibility reviews, including a lightweight extension to `schema.org` to support provenance and explainability (section 2)
- an implementation of the architecture demonstrating feasibility and value
- an evaluation on three datasets establishing state-of-the-art in one (Clef'18 CheckThat! Factuality task) and demonstrating capabilities and limitations of our approach (section 3)

[1]https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation

[2]https://credweb.org/signals-beta/

2 Linked Credibility Reviews

Linked Credibility Reviews (LCR) [Denaux and Perez-Gomez, 2020], is a (conceptual and data) model for composable and explainable misinformation detection. Our conceptual model defines a *Credibility Review*s (CR) and implements it as an extension of a `Review` as defined by the Schema.org vocabulary [Guha *et al.*, 2016][3]. We define a CR as a tuple $\langle d, r, c, p \rangle$, where the CR:

- reviews a *data item* d, this can be any linked-data node but will typically refer to articles, claims, websites, images, social media posts, people, publishers, etc.
- assigns a *credibility rating* r to the *data item* under review and qualifies it with a *rating confidence* c.
- *provides mandatory provenance information* p about:
  - *credibility signals* used to derive the credibility rating, which can be either (i) CRs for data items relevant to the data item under review or (ii) *ground credibility signals* (GCS), resources (which are not CRs) in databases curated by a trusted party.
  - the *author* of the review. The author can be a person, organization or bot. Bots are automated agents that produce CRs for supported data items based on a variety of strategies.

A key insight is that calculations of credibility are ultimately subjective and have to be modeled accordingly. Hence, the credibility rating provides a subjective (from the point-of-view of the author) measure of how much the credibility signals support or refute the content in data item. Provenance information is therefore crucial as it allows humans — e.g. end-users, bot developers— to retrace the CRs back to the ground credibility signals and assess the accuracy of the (possibly long) chain of bots (and ultimately humans) that were involved in reviewing the initial data item. The provenance also enables the generation of explanations for each step of the credibility review chain in a composable manner as each bot (or author) can describe its own strategy to derive the credibility rating based on the used credibility signals.

2.1 Reviewing Strategies and Implementations

In our ISWC paper [Denaux and Perez-Gomez, 2020], we formally define strategies for computing CRs and thus implementing bots that can collaborate in the construction of chained credibility reviews. Intuitively, the identified strategies were:

- **ground credibility signal lookup** from some trusted source. CR bots map simple queries to ground credibility signals. We demonstrated this by implementing two bots. The first one returns `ClaimReview` instances for known claims[4]. Our implementation is based on a database of 45K `ClaimReviews`. A second bot was derived by writing a wrapper around the existing MisinfoMe aggregation service [Mensio and Alani, 2019b], which produces credibility values for websites from existing services and datasets like OpenSources[6].
- **linking** some web content to review d with n other data items d'_i *of the same type*, for which a CR is available. In these cases, the polar similarity (e.g. similar but disagreeing) between contents have to be taken into account when propagating the credibility scores (and their confidences). We implemented a linking bot by combining two RoBERTa deep learning models for semantic similarity and stance detection.
- **decomposing** whereby the bot identifies relevant parts d'_i of the item-to-review d and requests CRs for those parts $CR_{d'_i}$. Like the linking bots, these require deriving new credibility ratings CR_{d_i} and confidences based on the relation between the whole and the parts. We implemented a decomposing bot using a proprietary (but common) sentence detection system optionally filtered using a RoBERTa checkworthiness checker.

2.2 Representing and Aggregating Ratings

For ease of computation, we recommend to represent credibility ratings and their confidences as follows:

- $r \in \Re$, must be in the range of $[-1.0, 1.0]$ where -1.0 means not credible and 1.0 means credible
- $c \in \Re$, must be in the range of $[0.0, 1.0]$ where 0.0 means no confidence at all and 1.0 means full confidence in the accuracy of r, based on the available evidence in p.

This representations makes it possible to define generic, relatively straightforward aggregation functions like selecting sub-reviews with the highest confidence value or the lowest credibility rating.

2.3 Extending `schema.org` for LCR

A large proportion of content online is already described by webmasters using `schema.org`, which already provides a vocabulary to describe `Reviews`. We identified some basic extensions compliant with the original definitions which allow us to contribute to the schema.org ecosystem. Mainly, we added various new subtypes of Reviews, a `Sentence` type, a `confidence` property to `schema:Rating` and we allow `Bots` to also author `CreativeWorks`. An overview of the `schema.org` based data model and our extensions is depicted in figure 1.

Figure 2 shows a diagram depicting how the various CR bots compose and collaborate to review a tweet. Example reviews generated by out bots are presented in table 1. Our implementation, `acred`, is available on GitHub[7].

3 Evaluation

We evaluated `acred` on three different datasets: first the Clef'18 CheckThat! Factuality Task (English part) [Nakov *et al.*, 2018] for predicting whether a check-worthy claim is `true`, `half-true` or `false`. It contains 74 and 139 claims for training and testing respectively. The second

[3] See also https://schema.org/Review

[4] `ClaimReview` markup[5], is defined by schema.org and enables fact-checkers to publish their work as machine readable structured data.

[6] https://github.com/BigMcLargeHuge/opensources

[7] https://github.com/rdenaux/acred

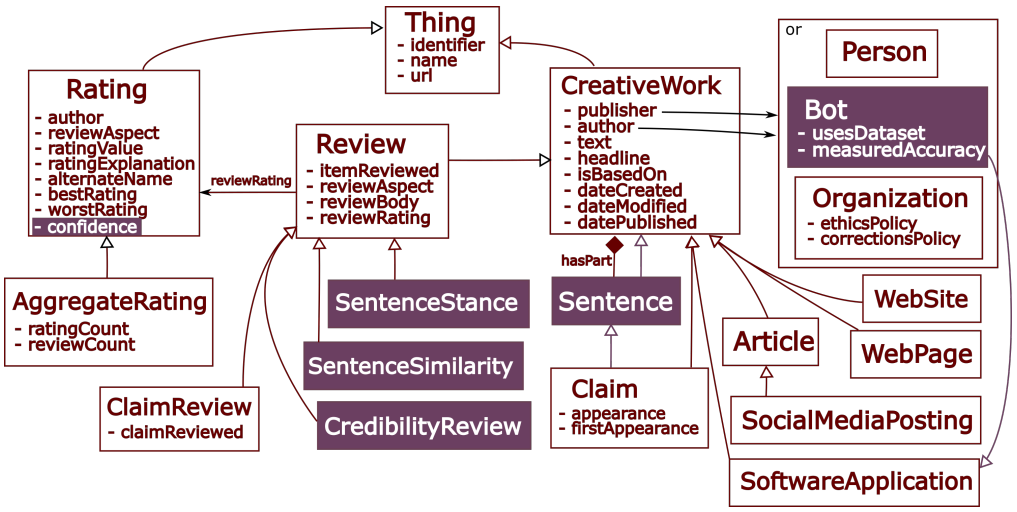

Figure 1: Linked Credibility Review data model, extending schema.org.

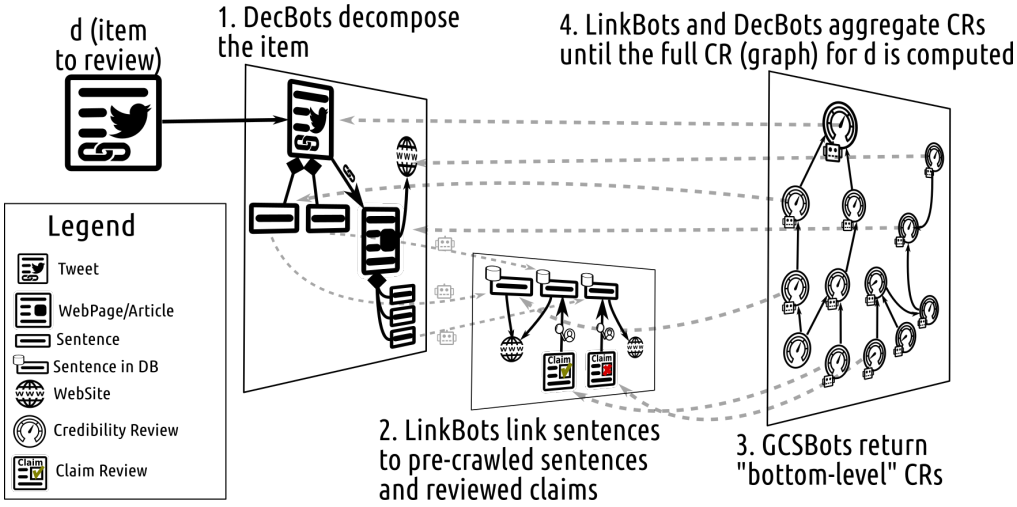

Figure 2: Depiction of acred bots collaborating to produce a CR for a tweet.

Bot	Example explanation
$\texttt{LookupBot}_{\texttt{ClaimRev}}$	Claim `'Ford is moving all of their small-car productin to Mexico.'` is *mostly not credible* based on a fact-check by politifact with normalised numeric ratingValue 2 in range [1-5]
$\texttt{LinkBot}^{\texttt{SemSim}}_{\texttt{Sentence}}$	Sentence `When Senator Clinton or President Clinton asserts that I said that the Republicans had had better economic policies since 1980, that is not the case.` seems *not credible* as it agrees with sentence: `Obama said that 'since 1992, the Republicans have had all the good ideas...'` that seems *not credible* based on a fact-check by politifact with textual rating 'false'. Take into account that the sentence appeared in site `www.cnn.com` that seems *credible* based on 2 review(s) by external rater(s) NewsGuard or Web Of Trust
$\texttt{DecBot}_{\texttt{SocMedia}}$	Sentence `Absolutely fantastic, there is know difference between the two facist socialist powers of today's EU in Brussels, and the yesteryears of Nazi Germany` in tweet agrees with: `'You see the Nazi platform from the early 1930s ... look at it compared to the (Democratic Party) platform of today, you're saying, 'Man, those things are awfully similar.''` that seems *not credible* based on a fact-check by politifact with textual claim-review rating 'false'"

Table 1: Example explanations generated by our bots.

label	r	c
credible	$r \geq 0.5$	$c > 0.7$
mostly credible	$0.5 > r \geq 0.25$	$c > 0.7$
uncertain	$0.25 > r \geq -0.25$	$c > 0.7$
mostly not credible	$-0.25 > r \geq -0.5$	$c > 0.7$
not credible	$-0.5 > r$	$c > 0.7$
not verifiable	any	$c \leq 0.7$

Table 2: Mapping of credibility `ratingValue` r and `confidence` c for coinform250.

system	MAE	F1	Macro AvgR
`acred`	0.6835	**0.4247**	0.4367
`acred`$^+$	**0.6475**	0.3741	0.4202
Copenhagen	0.7050	0.4008	**0.4502**
random	0.8345	0.3569	0.3589

Table 3: Results on `clef18` English test dataset compared to baselines . MAE (Mean Average Error) is the official metric in the competition.

dataset was the Politifact fragment of FakeNewsNet [Shu *et al.*, 2020] where the task is to predict whether articles are *fake* (420) or *real* (528). Finally, `coinform250`[8] is a dataset of 250 annotated tweets previously reviewed by fact-checkers and with associated `ClaimReviews` retrieved by MisinfoMe [Mensio and Alani, 2019a]. The original fact-checker labels were mapped onto six labels (see table 2) by 7 human raters achieving Fleiss κ 0.52 (moderate agreement). The fine-grained labels make this challenging but realistic.

For each dataset our prediction procedure consisted in steps to (i) read samples, (ii) convert them to the appropriate `schema.org` data items (`Sentence`, `Article` or `SocialMediaPost`), (iii) request a review from the appropriate acred CR bot and (iv) map the produced CR onto the dataset labels by defining confidence c and credibility rating r thresholds. For `clef18` we set $t = 0.75$, so that r values above t are `TRUE`, below $-t$ are `FALSE` and in between is `HALF-TRUE`. Table 2 shows thresholds for `coinform250`.

3.1 Results

Initial evauation on the datasets showed that `acred` was overly confident in some cases; this prompted us to introduce a modified version, `acred`$^+$ with custom heuristics to reduce the confidence and rating values for `Articles` when only website credibility signals are available and when stance was either "unrelated" or "discuss".

On `clef18`, `acred` established a new state-of-the-art result, which was further improved with `acred`$^+$ as shown in table 3. This result is noteworthy as the baseline systems, Copenhangen [Wang *et al.*, 2018] and random [Nakov *et al.*, 2018], used the training set of `clef18`; by contrast, `acred` did not use this data to finetune the underlying models.

On FakeNewsNet, `acred`$^+$ obtained state of the art results and `acred` obtained competitive results in line with strong baseline systems reported in the original paper [Shu *et al.*, 2020], shown in table 4. We consider as baselines systems

[8]https://github.com/co-inform/Datasets

System	Accuracy	Precision	Recall	F1
`acred`	0.586	0.499	**0.823**	0.622
`acred`$^+$	**0.716**	0.674	0.601	**0.713**
CNN	0.629	**0.807**	0.456	0.583
SAF/S	0.654	0.600	0.789	0.681

Table 4: Results on FakeNewsNet Politifact compared to baselines.

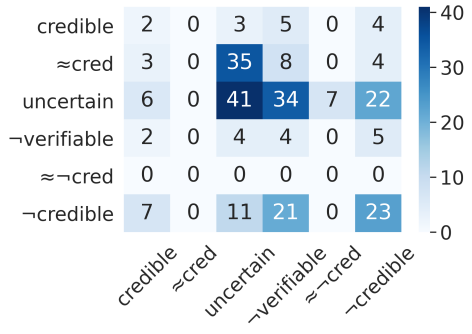

Figure 3: Confusion matrix `acred`$^+$ on `coinform250`. We use ≈ for *mostly* and ¬ for *not*. Rows are true labels, columns are predictions.

which only use the article content, since acred does not use social context signals yet. Baselines used 80% of the data for training and 20% for testing, while we used 100% for testing.

Finally, for the `coinform250` dataset, `acred`$^+$ obtains 0.279 accuracy which is well above a baseline of random predictions, which obtains 0.167 accuracy. The confusion matrix shown in figure 3 shows that `acred` tends to be overconfident in its predictions, while `acred`$^+$ is more cautious.

3.2 Discussion and Future Work

Our evaluation results show that our approach is capable of producing good results by integrating many readily available components. The proposed architecture initially does not require finetuning and can be reused to review a wide variety of web contents: single claims, tweets and articles; demonstrating composability and reusability. As a downside, errors in low-level modules (e.g. stance detection), may require handcrafting heuristic rules to correct. In the future it may be better to find automated ways to correct such errors. A caveat with our experimental results is that most datasets are (often indirectly) derived from fact-check articles and so is our database of claims (via `ClaimReviews`). This is bound to introduce noise, but lack of better datasets and evidence databases makes it hard to perform better experiments.

Our implementation can produce human understandable explanations and complex evidence graphs. We are researching how these outputs can be used to get user feedback to pinpoint the source of errors [Denaux *et al.*, 2020].

Acknowledgments

This work was supported by the European Comission under grant 770302 – Co-Inform – as part of the Horizon 2020 research and innovation programme.

References

[Babakar and Moy, 2016] Mevan Babakar and Will Moy. The State of Automated Factchecking. https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf, August 2016. Accessed: 2021-06-26.

[Cazalens *et al.*, 2018] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. A Content Management Perspective on Fact-Checking. In *The Web Conference*, 2018.

[Denaux and Perez-Gomez, 2020] Ronald Denaux and Jose Manuel Perez-Gomez. Linked Credibility Reviews for Explainable Misinformation Detection. In *19th International Semantic Web Conference*, nov 2020.

[Denaux *et al.*, 2020] Ronald Denaux, Flavio Merenda, and Jose Manuel Perez-Gomez. Towards Crowdsourcing Tasks for Accurate Misinformation Detection. In *Semantics for Online Misinformation Detection, Monitoring, and Prediciton. SEMIFORM@ISWC20*, volume 2722. CEUR-WS, nov 2020.

[Guha *et al.*, 2016] Ramanathan V Guha, Dan Brickley, and Steve Macbeth. Schema. org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016.

[Hassan *et al.*, 2017] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. Claim buster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2017.

[Marwick and Lewis, 2017] Alice Marwick and Rebecca Lewis. Media manipulation and disinformation online. *New York: Data & Society Research Institute*, 2017.

[Mensio and Alani, 2019a] Martino Mensio and Harith Alani. MisinfoMe: Who's Interacting with Misinformation? In *18th International Semantic Web Conference: Posters & Demonstrations*, 2019.

[Mensio and Alani, 2019b] Martino Mensio and Harith Alani. News Source Credibility in the Eyes of Different Assessors. In *Conference for Truth and Trust Online*. In Press, 2019.

[Nakov *et al.*, 2018] Preslav Nakov, Alberto Barrón-Cedeño, Reem Suwaileh, Lluís M ' Arquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, Giovanni Da, and San Martino. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372—-387, 2018.

[Papadopoulos *et al.*, 2016] Symeon Papadopoulos, Kalina Bontcheva, Eva Jaho, Mihai Lupu, and Carlos Castillo. Overview of the special issue on trust and veracity of information in social media. *ACM Transactions on Information Systems (TOIS)*, 34(3):1–5, 2016.

[Pérez-Rosas *et al.*, 2018] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic Detection of Fake News. In *COLING*, 2018.

[Pomerleau and Rao, 2017] Dean Pomerleau and Delip Rao. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. http://www.fakenewschallenge.org/, 2017. Accessed: 2021-06-26.

[Schiller *et al.*, 2021] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Stance Detection Benchmark: How Robust Is Your Stance Detection? *KI-Künstliche Intelligenz*, March 2021. https://doi.org/10.1007/s13218-021-00714-w.

[Shu *et al.*, 2020] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. *Big Data*, 8:171–188, 2020.

[Wang *et al.*, 2018] Dongsheng Wang, Jakob Grue Simonsen, Birger Larsen, and Christina Lioma. The Copenhagen Team Participation in the Factuality Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 Fact Checking Lab. *CLEF (Working Notes)*, 2125, 2018.

[Zubiaga *et al.*, 2018] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2), 2018.